# Challenges in Data Intensive Computing

**Jim Tomkins**

**Presented at SOS10**

**Maui, Hawaii**

**March 6 - 9, 2006**

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

NNSA National Nuclear Security Administration

Sandia National Laboratories

# Models of Interaction with HEC Storage Systems: HEC Machines

- Capability Scientific Computing - Thousands of Processors working on a single application with one to a few applications sharing the machine
  - Large Parallel Files - Restart, Graphics
  - Large load on meta-data services
  - Very large file system needed to support capability applications
  - High Bandwidth is needed from a single file system - Disk I/O is a system performance bottleneck
  - Write Dominated (~90% of disk I/O)
  - Defensive I/O increases stress on file system

# Models of Interaction with HEC Storage Systems: HEC Machines

- **Capacity Scientific Computing - Up to a thousand or so processors working on a single application with several to many applications sharing the machine**
  - **Small to Medium Size Parallel Files - Restart, Graphics**
  - **Multiple smaller file systems can make sense**
  - **Meta-data services can be distributed among file systems each with own meta-data support**
  - **Total Bandwidth needed is high but it can be divided among several file systems - Disk I/O is still a system performance bottleneck but less so than for capability computing**
  - **Write Dominated (~90% of disk I/O)**
  - **Defensive I/O increases stress on file system**

# Models of Interaction with HEC Storage Systems: At the Desktop or other On-Site Machines

- **Move selected data off HEC machine**
  - **Archival storage**
  - **Graphics processing**
- **Raw data is not moved from HEC machine to desktop**
  - **Display graphical images at desktop**
  - **HEC systems generate so much data that it would be impossible to move any significant amount of it to a desktop system even if the bandwidth were available.**

# Models of Interaction with HEC Storage Systems: From Remote Systems

- **Minimize the movement of raw data to remote systems.**
  - **Display graphical images remotely**
  - **Leave as much of the data at the site where it is generated as possible**
  - **Remote bandwidth is expensive - currently it is impractical to move significant amounts of data to remote locations.**

# Globally Accessible File System and HEC Systems

- **On-Site Globally Accessible File System**
  - **For all current large HEC systems the file system is a serious bottleneck.**
  - **A globally accessible file system will increase contention and reduce the efficiency of the HEC system.**
  - **The globally accessible model makes sense for archiving and visualization on-site but not for HEC systems.**

# Globally Accessible File System and HEC Systems

- **Remote Globally Accessible File System**
  - **Latency - locking, time-outs, packet size**
  - **Reliability - lost packets, data errors, network dropout**
  - **Performance - remote network bandwidth**

# Technical Challenges for a Globally Accessible File System

- Latency
  - In relative terms latency is increasing - distance and intervening electronics
  - Impact on HEC system performance for accessing remote data - wait time
- Bandwidth
  - On-site bandwidth is much less than HEC system bandwidth
  - Remote bandwidth is very expensive and will be no more than on-site bandwidth
  - Encryption adds to bandwidth issues
- Reliability
  - Number of switches
  - Cables
- Bottom Line - How to keep disk I/O from being a performance bottleneck for HEC systems.

# File System Development: What Should the Focus Be?

- **Parallel File Systems for HEC**
  - **Scalability - Bandwidth, Meta Data Services**
  - **Reliability - Meta-data, OST failover, disk rebuilds**
- **Data Movement Between Systems**
  - **Parallel data movers**
  - **Reliability**

# Conclusions

- **Globally accessible file systems don't make sense for an HEC system's direct disk I/O.**
  - **HEC system waiting on desktop or small server systems**
  - **Increased latency for storage that is farther from the HEC machine**
  - **Bandwidth will be an even greater issue**
  - **Sharing HEC system's storage with other systems will have a negative impact on HEC system performance. A 10K node machine could be forced to wait on a single desktop system for file access.**
- **Globally accessible file systems make sense for the situation where all the client systems are similar and disk I/O is not a mjor performance bottleneck.**